PROGRAM MATERIALS
Program #2910
January 15, 2019

Issues Facing In-House Counsel Working on Autonomous Vehicles

Celesq® AttorneysEd Center
www.celesq.com

5301 North Federal Highway, Suite 180, Boca Raton, FL 33487
Phone 561-241-1919 Fax 561-241-1969

JONES DAY | One Firm Worldwide℠

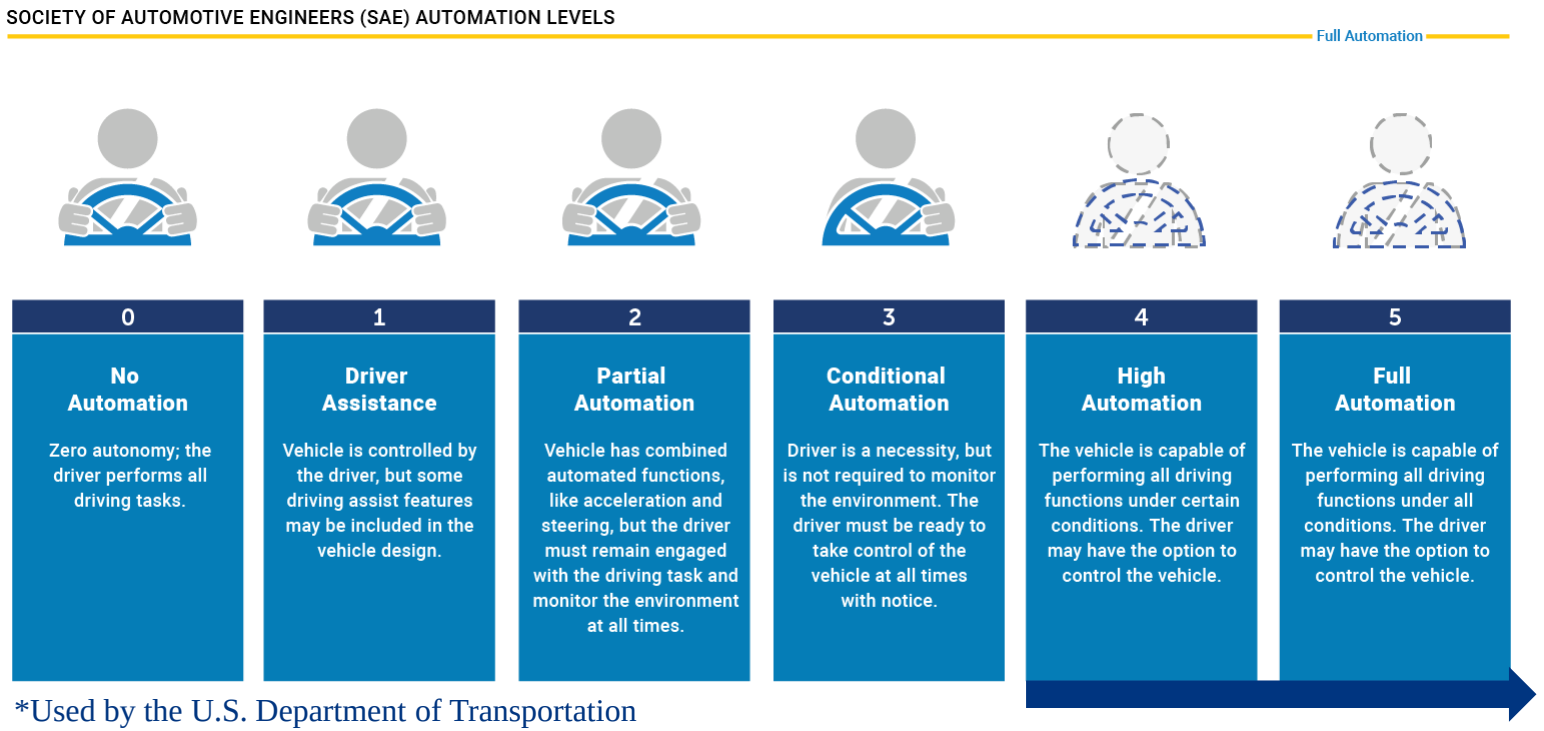
Autonomous – Vehicles

Considerations for In-House Attorneys

Jeffrey J. Jones
Jones Day Detroit
jjjones@jonesday.com
313-230-7950

NAI-1505296924v1

Working Definitions – Automated Vehicles



*Used by the U.S. Department of Transportation

Working Definitions – Artificial Intelligence

- Human intelligence involves:
  - Ability to analyze data collected by the senses
  - In light of prior experience
  - To reach conclusions and make decisions
  - To learn from experience
- Artificial intelligence is the ability of a computing device to perform these functions

Working Definitions – Machine Learning

- A common method of creating artificial intelligence
- Allows a computing device to obtain and apply knowledge without being explicitly programmed
- Machine learning, for example, involves training a model to learn to recognize objects within images – based on feeding large amounts of data containing known objects within images.

Working Definitions – Deep Learning

- One type of machine learning
- Based on structure and function of human brain – the interconnection of neurons
- Neural network – layers of connected nodes
- Input notes
- Performance nodes recognize
    - Shapes
    - Features
    - Images
- Output nodes

Uses of Artificial Intelligence in AVs

- Sensing entities around the vehicle, what they are doing, and where they are going
- Mapping the roadway and environment around the vehicle
- Negotiating traffic conditions
- Predictive maintenance
- In-user experience / individualized marketing
- Automotive insurance risk assessment

Legal and Ethical Issues Raised by Use of Artificial Intelligence in AVs

- Liability – when autonomous systems fail
- Bias – systems trained on data curated by humans creates potential for implicit bias
- Privacy – use / disclosure of personal data
- Cybersecurity – locking out bad actors

The Coming Collision Between Autonomous Vehicles and the Liability System

"Because drivers are found to be at fault in a large majority of current automobile accidents, removing the driver from the liability equation in autonomous vehicles will have important implications. Of course, by removing driver error as a factor, the frequency of accidents should go down, which is one of the key potential benefits of an autonomous vehicle in the first place. **When an accident does occur though, the vehicle manufacturer, or some other party involved in the design, manufacture, or operation of the autonomous vehicle is likely to be held liable for a higher proportion of the accidents.** This will be more likely to occur with autonomous vehicles than it currently does with conventional vehicles. In other words, when an autonomous vehicle does crash, most likely something went wrong with the collision avoidance system or the vehicle encountered conditions that it was not adequately programmed to address. Unlike a conventional vehicle crash, where the vehicle malfunction involves some sort of defect, such as a tire blowout or gas tank explosion, the malfunction in an autonomous vehicle will usually be a programming error or system failure that could implicate several different potentially liable parties."

-- Gary E. Marchant & Rachel A. Lindor, 52 *Santa Clara L. Rev.* 1327-28 (2012).

Who is liable?

- The list of potential parties includes:
  - Vehicle manufacturer
  - Manufacturer of a component used in the autonomous system
  - Software engineer who programmed the code for the autonomous operation of the vehicle
  - The road designer in the case of an intelligent road system that helps control the vehicle
  - The driver who “failed” to take control

If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Liability Claims

"Where liability is formally based on fault, manufacturers must take all reasonable steps proportionate to the risk to reduce foreseeable risks of harm. In most instances, this means that manufacturers must do the following: exercise reasonable care to eliminate all substantial dangers from their products that can reasonably be designed away; warn consumers about all substantial hidden dangers that remain; produce their products in such a way as to minimize dangerous manufacturing flaws; and be careful to avoid misrepresenting the safety of their products. In order to avoid negligence-based liability, manufacturers must exercise reasonable care in all of these respects; if they do so, responsibility for any remaining dangers in the use of the products – whether defects, inherent hazards, or generic risks – must be borne by consumers. In the context of strict liability, manufacturers of unavoidably hazardous products must ensure that their products are free of production defects and must warn consumers of hidden dangers."

-- Sabine Gless, Emily Silverman & Thomas Weigend, SSRN (2016)

Defending Product Liability

- Although the overall safety benefit of AVs will not provide a complete liability shield, manufacturers of AVs could try to defend their vehicle by demonstrating that it is safer overall than the conventional vehicle it replaces.
    - However, the plaintiff's expert may claim (with the benefit of hindsight) that the manufacturer should have known about and adopted an alternative safer design.
- By fully disclosing the potential risks of the vehicle, including the likely failure modes and some approximate sense of their probability, the manufacturer paves the way for an assumption of risk defense.
- As AVs become commonplace, manufacturers may also argue that they are part of the “normal” risks of life, comparable to lightning or falling trees. The self-driving car will not be liable for its generally foreseeable malfunctioning but only for harm incurred due to preventable construction, programming, or operating errors.

Will Knight: "The Dark Secret at the Heart of AI"

Machine learning for autonomous cars:

"Information from the vehicles' sensors goes straight into a huge network of artificial neurons that process the data and then deliver the commands required to operate the steering wheel, the brakes and other systems. The result seems to match the responses you'd expect to see from a human driver. But what if one day it did something unexpected – crashed into a tree, or sat at a green light? As things stand now, it might be difficult to find out why. The system is so complicated that even the engineers who designed it may struggle to isolate the reason for any single action. And you can't ask it: there is no obvious way to design such a system so that it could always explain why it did what it did."

Developing Ethical AVs

- Developing ethical AVs requires communication and planning around:
  - Transparency
  - Accountability
  - Privacy
  - Cybersecurity
- The Trolley Problem isn’t about finding perfect solutions but assessing and establishing public expectations
- How decisions are made are as important as the decisions themselves and impact liability

Why Some May Not Trust AVs

- Mistakes by creators, designers, programmers in underlying AI
    - Rely on inaccurate or incomplete data
    - Failure to recognize own bias
    - Miscalculation
- Fear of hacking or manipulation
- Misunderstanding of or objection to the vehicle's decision-making functions / moral framework
- Uncertainty surrounding liability

What Can be Done?

Build Transparency and Accountability

- A key prerequisite to the public's acceptance of automated vehicles (and their underlying AI) in performing key driving functions is trust.
- Ingredients to building trust include:
  - Transparency
    - U.S. Department of Transportation's *AV 3.0* encourages companies developing, testing, and deploying automation technology to be transparent about how safety is being achieved
  - Accountability
    - Institute of Electrical and Electronics Engineers (IEEE) proposes "ethically aligned design" to help prove why systems operate in certain ways to:
      - address legal issues of culpability
      - avoid confusion or fear within the general public

TRANSPARENCY

- How are decisions made?
- Different needs for different stakeholders:
  - Users
  - Regulatory bodies
  - Advocates
  - Society at large
- A basis for informed decisions / consent
- Balance with protecting the proprietary information that underlies the technology

ACCOUNTABILITY

- Why was that decision made?
- Key principles:
  - Responsibility
  - Explainability
  - Accuracy
  - Auditability
  - Fairness
- Justifying actions
- Assigning liability

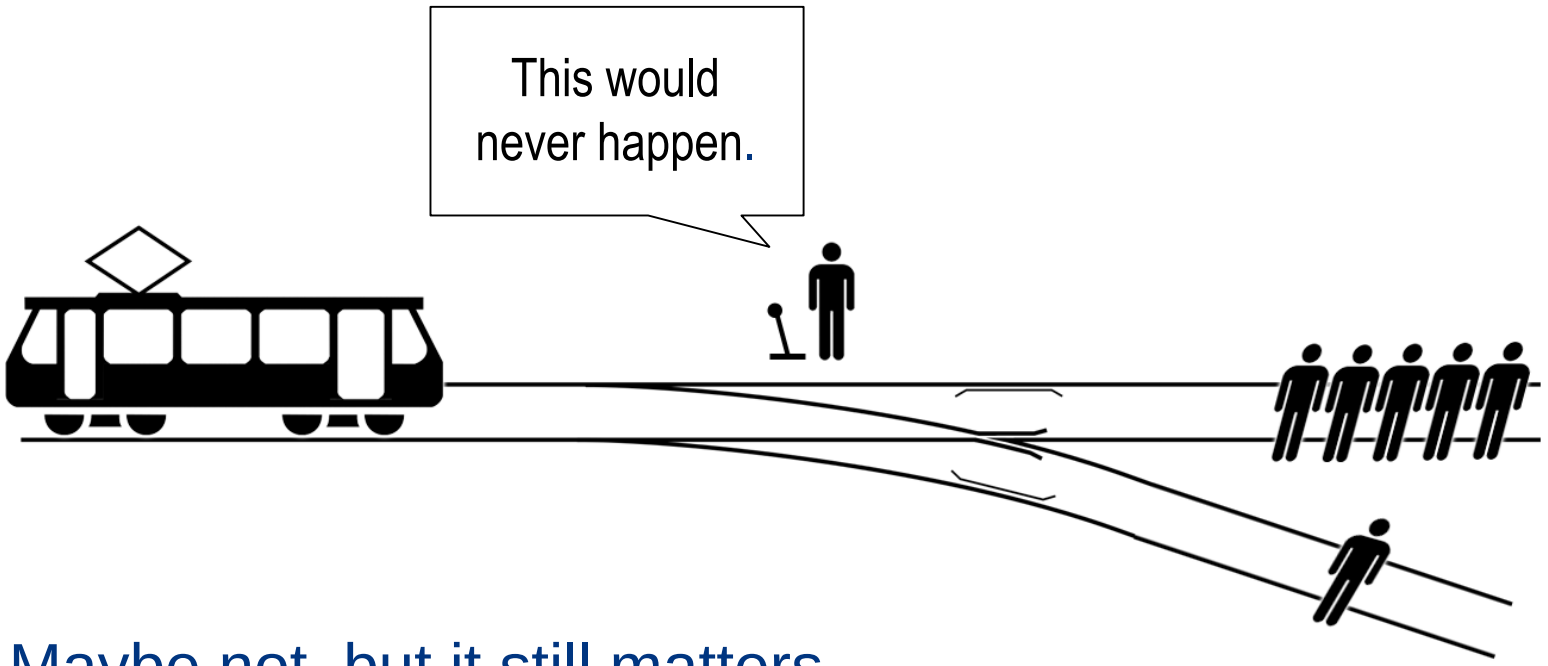
Manage Privacy

- Decide how vehicle data generated by road users will be forwarded, used, and secured
- Balance functional safety and informational self-determination
- Transparency around data collection and management practices
- Be aware of jurisdictional differences
- Informed consent from users

Maintain Cybersecurity

- Key principles of cyber security for connected and automated vehicles (UK 2017):
  - Organizational security is owned, governed, and promoted
  - Security risks are assessed and managed appropriately and proportionately, including those specific to the supply chain
  - Organizations need product aftercare and incident response to ensure systems are secure over their lifetime
  - All organizations, including sub-contractors, suppliers, and potential 3rd parties, work together to enhance the security of the system
  - Systems are designed using a defense-in-depth approach
  - The security of the software is managed throughout its lifetime
  - The storage and transmission of data is secure and can be controlled
  - The system is designed to be resilient to attacks and respond appropriately when its defenses or sensors fail

Address Dilemmas



Maybe not, but it still matters.

Reality of ethical questions?

Robot Cars And Fake Ethical Dilemmas

Patrick Lin Contributor ⓘ

Something feels dishonest about the moral panic over self-driving cars. It usually involves bizarre crash scenarios that would (probably) never happen in real life. Does it matter that the scenarios are artificial or unrealistic?

JONES DAY®

Robot Cars And Fake Ethical Dilemmas

"The provocative scenarios in these ethics discussions are known as *thought experiments*, which are like science experiments, except they're all conducted in our minds, without labs and equipment. Thought experiments are designed to be intuition pumps, probing the limits of what we believe to be true. They isolate and stress particular beliefs in artificial scenarios, because real-world scenarios are often too messy with uncontrolled, entangled variables. . . . The job of these thought experiments is to force us to think more carefully about ethical priorities, not to simulate reality. Our assumptions may work for the normal cases, but if technology developers don't understand the limits of those assumptions—if we don't know where they fracture and fail—then they arguably haven't done their *due diligence* in designing the product, because we can't expect them all to be normal cases. That is important in, say, a product-liability lawsuit against a manufacturer for a fatal crash caused by its automated car's programming. And the weirder the crash, the bigger the news headlines, and this can have a snowball effect on public perceptions, market share, and regulator responses."

-- Patrick Lin, Forbes.com (April 3, 2017)

Test Assumptions

- “Edge cases” draw out subtle assumptions that may underlie decisions in everyday scenarios
- Assumptions about “normal” behavior have limits
- Evaluating consumer appetite for risk, both ordinary and extraordinary, may help develop social consensus
- Perfect solutions may not be possible. Well-documented and thoughtful explanations offer defenses to unexpected outcomes.

Programming for Ethical AVs

- No need to reinvent the wheel. Use current law.
- There is a notion that smart machines need to be able to render moral decisions individually but this does not take into account the fact that what is “right” or “legal” is not necessarily subject to individual deliberations and choice.
- Self-driving cars will have to obey the same laws as human drivers – laws that have already been collectively agreed upon by society.
- Ethical decisions left to the individual are only those which society has already ruled, rightly or wrongly, are not significantly harmful. They remain unconstrained by regulation or attendant legislation.
- Individuals are free to make these decisions as they see fit until such time as society changes the extent and kind of behavior regulated through collective action (i.e., by law and law enforcement).

Programming for Ethical AVs (continued)

- Where the law is silent and individuals are free to make decisions as they see fit, ethics bots can determine user preferences and guide decision-making accordingly.
- Just as companies mine data to discover consumer preferences, ethics bots may be used to extract specific ethical preferences from a user and subsequently apply these preferences to operations of the user's machine.
- Many of the ethical decisions that smart machines are said to have to make, need not and should not be made by them because they are already addressed by law. These choices are made for the machines by society, using legislatures and courts. Many of the remaining ethical decisions can be made by ethics bots, which align the cars' "conduct" with the moral preferences of their owners.
- There will always be some incidents that cannot be foreseen and programmed – the "edge cases." If these situations are repeated, humans can address them through new laws or updates to the ethics bot.

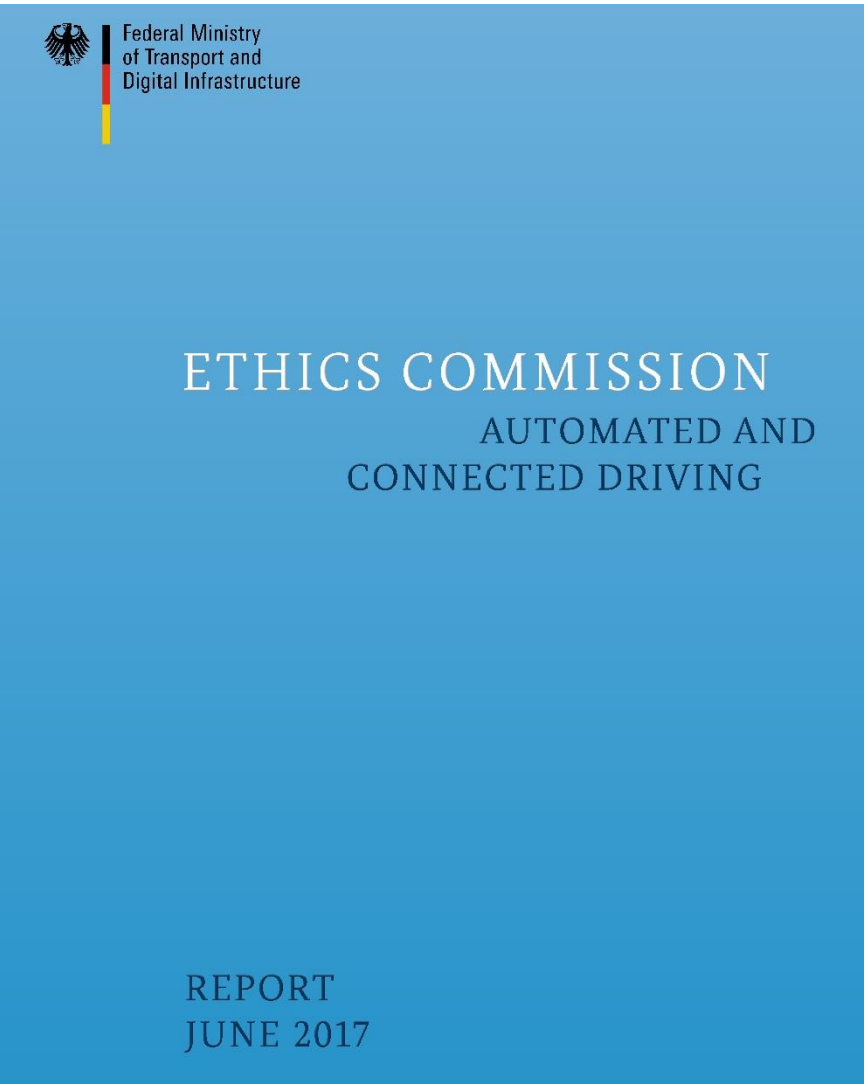
Programming for Ethical AVs (continued)

- In the meantime:
  - Society can set limits (e.g., law and regulations);
  - Users can provide ethical guidance in other situations; and
  - Models can be developed to compensate those harmed if and when the edge cases occur.

Other Programming Approaches

- **Top-down**: Ethical principles are programmed *directly* into the car's guidance system (e.g., Asimov's Three Laws of Robotics, the Ten Commandments, or a general moral philosophy such as utilitarianism). Instead of programming specific responses, the car makes choices based on whatever moral philosophy has been implanted into its AI program. However, there is always the possibility that, at some point or another, adherence to a particular philosophy will lead to actions and outcomes that others will find morally unacceptable.
- **Bottom-up**: The machines are programed to learn to render ethical decisions by *observing* human behavior in actual situations without any formal rules or particular moral philosophy. However, there are time constraints on this approach. It may take several lifetimes to come across all of the issues a single driver would ever face. Data may be aggregated across millions of users, but this raises the possibility that instead of teaching cars to be ethical, we teach them to do what is common (which could be the instinctive, rather than thoughtful, actions of humans).

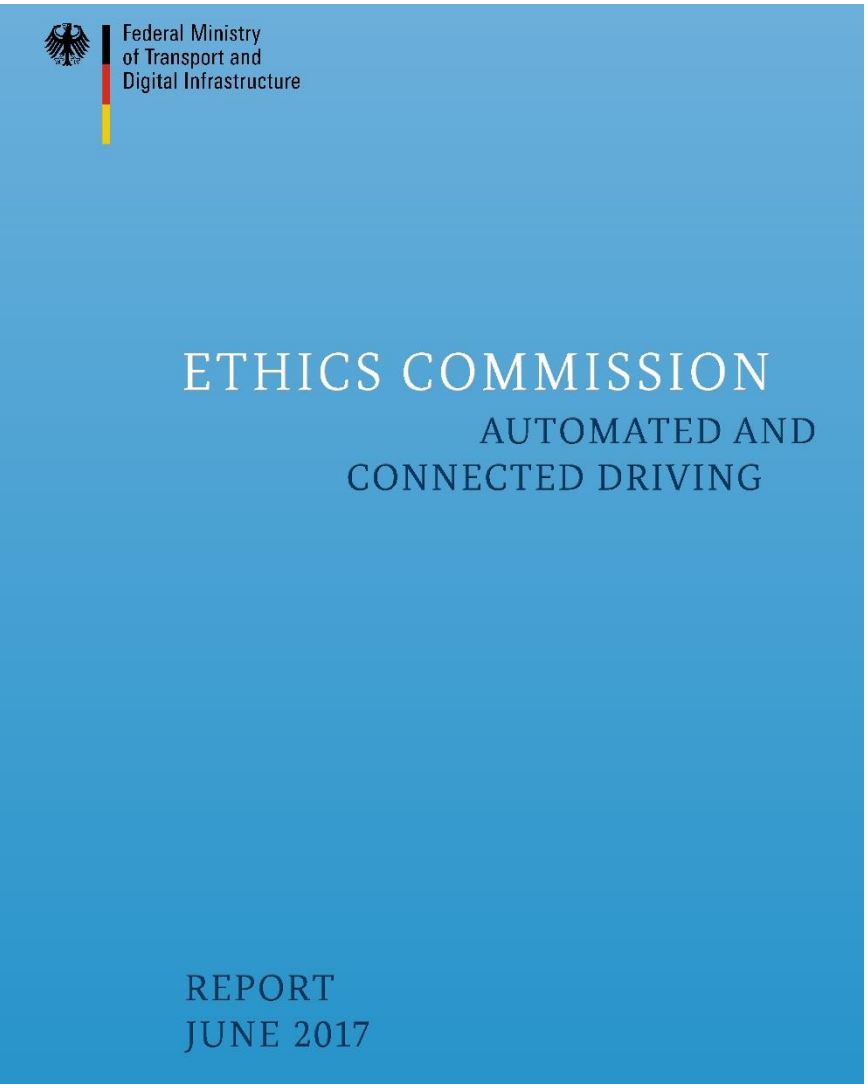
Proposed ethical rules for automated and connected vehicular traffic



Proposed Rule No. 1

The Primary purpose of partly and fully automated transport systems is to improve safety for all road users. Another purpose is to increase mobility opportunities and to make further benefits possible. Technological development obeys the principle of personal autonomy, which means that individuals enjoy freedom of action for which they themselves are responsible.

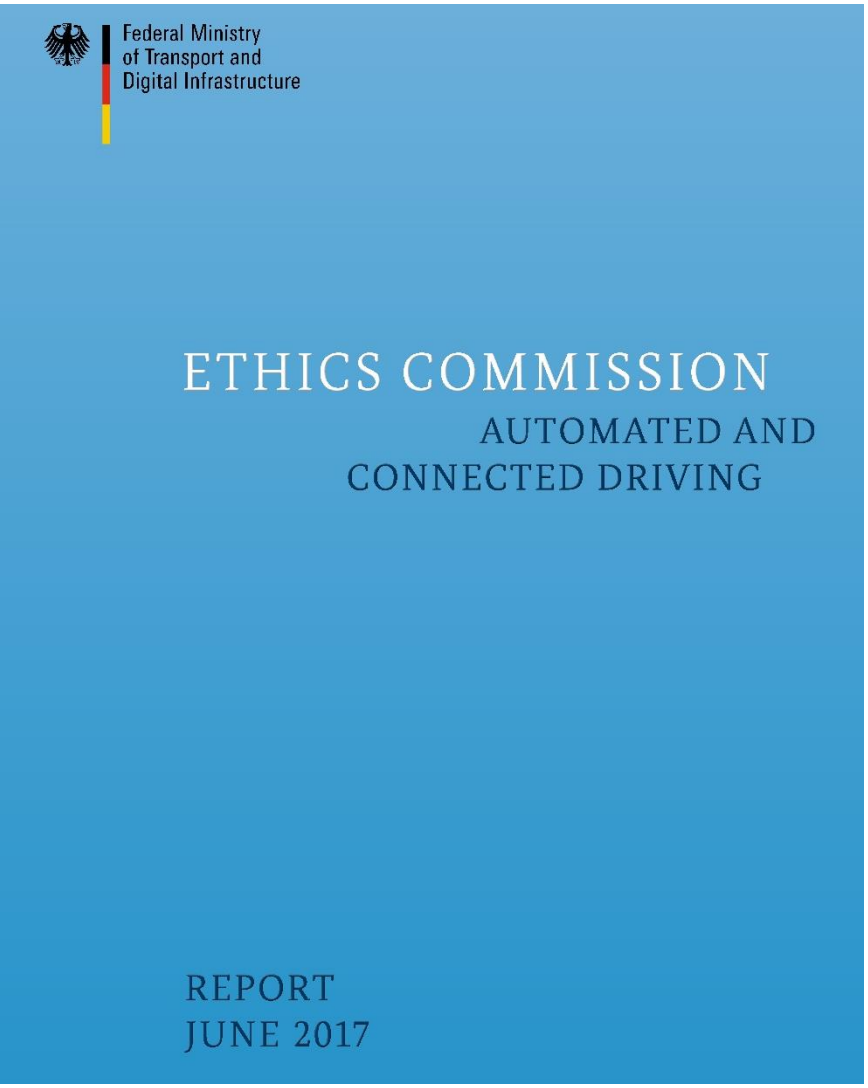
JONES DAY®



Proposed Rule No. 2

The protection of individuals takes precedence over all other utilitarian considerations. The objective is to reduce the level of harm until it is completely prevented. The licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving, in other words a positive balance of risks.

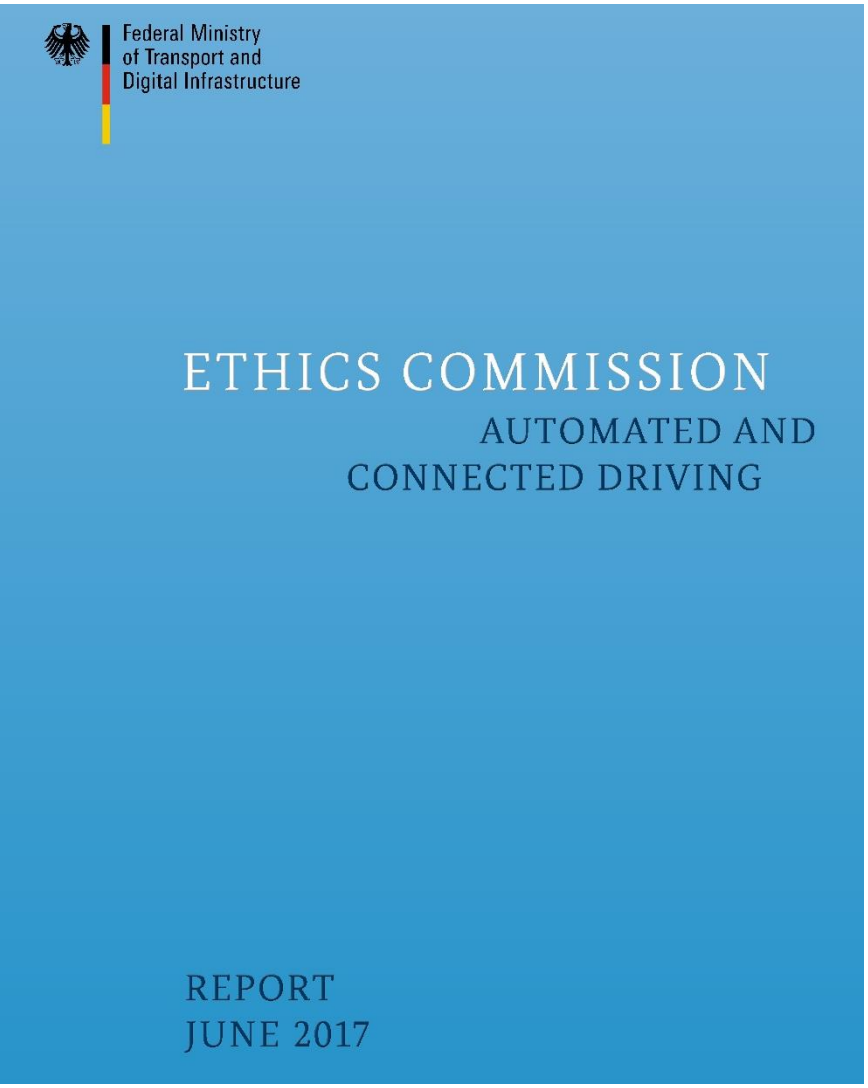
JONES DAY®



Proposed Rule No. 3

The public sector is responsible for guaranteeing the safety of the automated and connected systems introduced and licensed in the public street environment. Driving systems thus need official licensing and monitoring. The guiding principle is the avoidance of accidents, although technologically unavoidable residual risks do not militate against the introduction of automated driving if the balance of risks is fundamentally positive.

JONES DAY®



Proposed Rule No. 4

- The personal responsibility of individuals for taking decisions is an expression of a society centered on individual human beings, with their entitlement to personal development and their need for protection. The purpose of all governmental and political regulatory decisions is thus to promote the free development and the protection of individuals. In a free society, the way in which technology is statutorily fleshed out is such that a balance is struck between maximum personal freedom of choice in a general regime of development and the freedom of others and their safety.

Proposed Rule No. 5

- Automated and connected technology should prevent accidents whenever this is practically possible. Based on the state of the art, the technology must be designed in such a way that critical situations do not arise in the first place. These include dilemma situations, in other words a situation in which an automated vehicle has to "decide" which of two evils, between which there can be no trade-off, it necessarily has to perform. In this context, the entire spectrum of technological options – for instance from limiting the scope of application to controllable traffic environments, vehicle sensors and braking performance, signals for persons at risk, right up to preventing hazards by means of "intelligent" road infrastructure – should be used and continuously evolved. The significant enhancement of road safety is the objective of development and regulation, starting with the design and programming of the vehicles such that they drive in a defensive and anticipatory manner, posing as little risk as possible to vulnerable road users.

Proposed Rule No. 6

- The introduction of more highly automated driving systems, especially with the option of automated collision prevention, may be socially and ethically mandated it if can unlock existing potential for damage limitation. Conversely, a statutorily imposed obligation to use fully automated transport systems or the causation of practical inescapability is ethically questionable if it entails submission to technological imperatives (prohibition on degrading the subject to a mere network element).

Proposed Rule No. 7

- In hazardous situations that prove to be unavoidable, despite all technological precautions being taken, the protection of human life enjoys top priority in a balancing of legally protected interests. Thus, within the constraints of what is technologically feasible, the systems must be programmed to accept damage to animals or property in a conflict if this means that personal injury can be prevented.

Proposed Rule No. 8

- Genuine dilemmatic decisions, such as a decision between one human life and another, depend on the actual specific situation, incorporating “unpredictable” behavior by parties affected. They can thus not be clearly standardized, nor can they be programmed such that they are ethically unquestionable. Technological systems must be designed to avoid accidents. However, they cannot be standardized to complex or intuitive assessments of the impacts of an accident in such a way that they can replace or anticipate the decision of a responsible driver with the moral capacity to make correct judgements. It is true that a human driver would be acting unlawfully if he killed a person in an emergency to save the lives of one or more other persons, but he would not necessarily be acting culpably. Such legal judgements, made in retrospect and taking special circumstances into account, cannot readily be transformed into abstract/general ex ante appraisals and thus also not into corresponding programming activities. For this reason, perhaps more than any other, it would be desirable for an independent public sector agency (for instance a Federal Bureau for the Investigation of Accidents Involving Automated Transport Systems or a Federal Office for Safety in Automated and Connected Transport) to systematically process the lessons learned.

Proposed Rule No. 9

- In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties.

Proposed Rule No. 10

- In the case of automated and connected driving systems, the accountability that was previously the sole preserve of the individual shifts from the motorist to the manufacturers and operators of the technological systems and to the bodies responsible for taking infrastructure, policy and legal decisions. Statutory liability regimes and their fleshing out in the everyday decisions taken by the courts must sufficiently reflect this transition.

Proposed Rule No. 11

- Liability for damage caused by activated automated driving systems is governed by the same principles as in other product liability. From this, it follows that manufacturers or operators are obliged to continuously optimize their systems and also to observe systems they have already delivered and to improve them where this is technologically possible and reasonable.

Proposed Rule No. 12

- The public is entitled to be informed about new technologies and their deployment in a sufficiently differentiated manner. For the practical implementation of the principles developed here, guidance for the deployment and programming of automated vehicles should be derived in a form that is as transparent as possible, communicated in public and reviewed by a professionally suitable independent body.

Proposed Rule No. 13

- It is not possible to state today whether, in the future, it will be possible and expedient to have the complete connectivity and central control of all motor vehicles within the context of a digital transport infrastructure, similar to that in the rail and air transport sectors. The complete connectivity and central control of all motor vehicles within the context of a digital transport infrastructure is ethically questionable if, and to the extent that, it is unable to safely rule out the total surveillance of road users and manipulation of vehicle control.

Proposed Rule No. 14

- Automated driving is justifiable only to the extent to which conceivable attacks, in particular manipulation of the IT system or innate system weaknesses, do not result in such harm as to lastingly shatter people's confidence in road transport.

Proposed Rule No. 15

- Permitted business models that avail themselves of the data that are generated by automated and connected driving and that are significant or insignificant to vehicle control come up against their limitations in the autonomy and data sovereignty of road users. It is the vehicle keepers and vehicle users who decide whether their vehicle data that are generated are to be forwarded and used. The voluntary nature of such data disclosure presupposes the existence of serious alternatives and practicability. Action should be taken at an early state to counter a normative force of the factual, such as that prevailing in the case of data access by the operators of search engines or social networks.

Proposed Rule No. 16

- It must be possible to clearly distinguish whether a driverless system is being used or whether a driver retains accountability with the option of overruling the system. In the case of non-driverless systems, the human-machine interface must be designed such that at any time it is clearly regulated and apparent on which side the individual responsibilities lie, especially the responsibility for control. The distribution of responsibilities (and thus of accountability), for instance with regard to the time and access arrangements, should be documented and stored. This applies especially to the human-to-technology handover procedures. International standardization of the handover procedures and their documentation (logging) is to be sought in order to ensure the compatibility of the logging or documentation obligations as automotive and digital technologies increasingly cross national borders.

Proposed Rule No. 17

- The software and technology in highly automated vehicles must be designed such that the need for an abrupt handover of control to the driver (“emergency”) is virtually obviated. To enable efficient, reliable and secure human-machine communication and prevent overload, the system must adapt more to human communicative behavior rather than requiring humans to enhance their adaptive capabilities.

Proposed Rule No. 18

- Learning systems that are self-learning in vehicle operation and their connection to central scenario databases may be ethically allowed if, and to the extent that, they generate safety gains. Self-learning systems must not be deployed unless they meet the safety requirements regarding functions relevant to vehicle control and do not undermine the rules established here. It would appear advisable to hand over relevant scenarios to a central scenario catalogue at a neutral body in order to develop appropriate universal standards, including any acceptance tests.

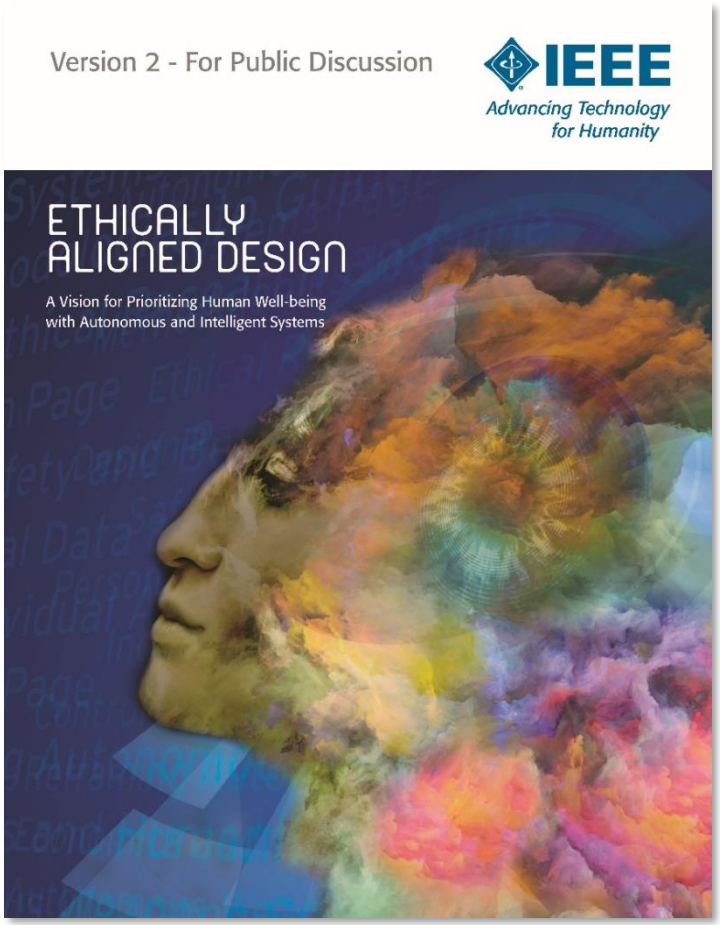
Proposed Rule No. 19

- In emergency situations, the vehicle must autonomously, i.e. without human assistance, enter into a “safe condition”. Harmonization, especially of the definition of a safe condition or of the handover routines, is desirable.

Proposed Rule No. 20

- The proper use of the automated systems should form part of people's general digital education. The proper handling of automated driving systems should be taught in an appropriate manner during driving tuition and tested.

Does the IEEE Have Any Proposals?



JONES DAY®

IEEE Initiative on Ethics of Autonomous Intelligent Systems (“AIS”)

- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (“The IEEE Global Initiative”) is a program of The Institute of Electrical and Electronics Engineers (“IEEE”).

Mission of IEEE Global Initiative on Ethics of AIS

"Our goal is that Ethically Aligned Design will provide insights and recommendations that provide a key reference for the work of technologists in the related fields of science and technology in the coming years. To achieve this goal, in the current version of Ethically Aligned Design (EAD2v2), we identify pertinent "Issues" and "Candidate Recommendations" **we hope will facilitate the emergence of national and global policies that align with these principles."**

A second goal of The IEEE Global Initiative is to provide recommendations for IEEE Standards based on Ethically Aligned Design.

IEEE Working Groups

IEEE P7000™ - Model Process for Addressing Ethical Concerns During System Design

IEEE P7001™ - Transparency of Autonomous Systems

IEEE P7002™ - Data Privacy Process

IEEE P7003™ - Algorithmic Bias Considerations

IEEE P7004™ - Standard on Child and Student Data Governance

IEEE P7005™ - Standard for Transparent Employer Data Governance

IEEE P7006™ - Standard for Personal Data Artificial Intelligence (AI) Agent

IEEE P7007™ - Ontological Standard for Ethically Driven Robotics and Automation Systems

IEEE P7008™ - Standard for Ethically Driven Nudging for Robotic, Intelligent, and Automation Systems

IEEE P7009™ - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

IEEE P7010™ - Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems

General Principles of IEEE Global Initiative on Ethics of AIS

- **Human Rights:** Ensure they do not infringe on internationally recognized human rights
- **Well-being:** Prioritize metrics of well-being in their design and use
- **Accountability:** Ensure that their designers and operators are responsible and accountable
- **Transparency:** Ensure they operate in a transparent manner
- **Awareness of misuse:** Minimize the risks of their misuse

Transparency and Individual Rights (vs. Government) Objectives

- Parties, their lawyers, and courts must have **reasonable access to all data and information generated and used by such systems** employed by governments and other state authorities
- **The logic and rules embedded in the system must be available to overseers**, if possible, and subject to risk assessments and rigorous testing
- The systems should generate **audit trails** recording the facts and law supporting decisions and they should be amenable to third-party verification
- The general public should know who is making or supporting ethical decisions of such systems through investment

IEEE Recommendation for Transparency Generally

- Develop new standards* that describe **measurable, testable levels of transparency**, so that systems can be objectively assessed and levels of compliance determined. For designers, such standards will provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency. (The mechanisms by which transparency is provided will vary significantly, for instance **1) for users of care or domestic robots, a why-did-you-do-that button which, when pressed, causes the robot to explain the action it just took**, 2) for validation or certification agencies, the algorithms underlying the AIS and how they have been verified, and 3) for accident investigators, secure storage of sensor and internal state data, comparable to a flight data recorder or black box.)

*Note that IEEE Standards Working Group P7001™ has been set up in response to this recommendation.

IEEE on Accountability

- Manufacturers of AIS systems must be able to provide programmatic-level accountability proving why a system operates in certain ways to address legal issues of culpability and if necessary apportion culpability among several responsible designers, manufacturers, owners, and/or operators to avoid confusion or fear within the general public.

IEEE Recommendations for Accountability

1. Legislatures/courts should clarify issues of responsibility, culpability, liability, and accountability for AIS where possible during development and deployment (so that manufacturers and users understand their rights and obligations).
2. Designers and developers of AIS should remain aware of, and take into account when relevant, the diversity of existing cultural norms among the groups of users of these AIS.
3. Multi-stakeholder ecosystems should be developed to help create norms (which can mature to best practices and laws) where they do not exist because AIS-oriented technology and their impacts are too new.

IEEE Recommendations for Accountability

4. Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular AIS. Manufacturers/operators/owners of AIS should register key, high-level parameters, including:
    - Intended use
    - Training data/training environment (if applicable)
    - Sensors/real world data sources
    - Algorithms
    - Process graphs
    - User interfaces
    - Actuators/outputs

IEEE – Approaches to Embedding Values into AIS

A. Top-down approaches, where the system (e.g., a software agent) has some symbolic representation of its activity, and so can identify specific states, plans, or actions as ethical/unethical with respect to particular ethical requirements.

B. Bottom-up approaches, where the system (e.g., a learning component) builds up, through experience of what is to be considered ethical/unethical in certain situations, an implicit notion of ethical behavior.

Failures Will Occur – Mitigation Strategies

- A systematic risk analysis and management approach tries to anticipate potential points of failure (e.g., norm violations) and, where possible, develops some ways to mitigate or remove the effects of failures.
- Because not all risks and failures are predictable, especially in complex human-machine interactions in social contexts, additional mitigation mechanisms must be made available. Designers are strongly encouraged to augment the architectures of their systems with components that handle unanticipated norm violations with **a fail-safe, such as symbolic "gateway" agents**. The fail-safe components need to be extremely reliable and protected against security breaches.

IEEE – Evaluating the Implementation of AIS

- Evaluation by:
  - “First Parties” – Designers, Manufacturers, Users
  - “Third Parties” – Independent Testers, Regulators

IEEE – Evaluating the Implementation of AIS

- AIS should have sufficient transparency to allow evaluation by third parties, including regulators, consumer advocates, ethicists, post-accident investigators, or society at large. However, transparency can be severely limited in some systems, especially in those that rely on machine learning algorithms trained on large data sets. The data sets may not be accessible to evaluators; the algorithms may be proprietary information or mathematically so complex that they defy common-sense explanation; and even fellow software experts may be unable to verify reliability and efficacy of the final system because the system's specifications are opaque.
- For less inscrutable systems, numerous techniques are available to evaluate the implementation of an AIS's norm conformity. **On one side there is formal verification, which provides a mathematical proof that the AIS will always match specific normative and ethical requirements** (typically devised in a top-down approach). **This approach requires access to the decision-making process and the reasons for each decision.**

IEEE Recommendation – Documentation

- **Software engineers should be required to document all of their systems** and related data flows, their performance, limitations, and risks. Ethical values that have been prominent in the engineering processes should also be explicitly presented as well as empirical evidence of compliance and methodology used, such as data used to train the system, algorithms and components used, and results of behavior monitoring. **Criteria for such documentation could be: auditability, accessibility, meaningfulness, and readability.**

IEEE Recommendation – Limit Use of "Black Boxes"

- "Deep" machine learning processes are a growing source of "black-box" software. At least for the foreseeable future, AI developers will likely be unable to build systems that are guaranteed to operate exactly as intended or hoped for in every possible circumstance. Yet, the responsibility for resulting errors and harms remains with the humans that design, build, test, and employ these systems.

IEEE Recommendation – Limit Use of "Black Boxes"

- Technologists should be able to characterize what their algorithms or systems are going to do via transparent and traceable standards. To the degree possible, these characterizations should be predictive, but given the nature of AIS, they might need to be more retrospective and mitigation oriented. Such standards may include preferential adoption of effective design methodologies for building **"explainable AI" (XAI) systems** that can provide justifying reasons or other reliable "explanatory" data illuminating the cognitive processes leading to, and/or salient bases for, their conclusions.

IEEE Recommendation – Limit use of “Black Boxes”

- Software engineers should employ “black-box” (opaque) software services or components only with extraordinary caution and ethical care, as they tend to produce results that cannot be fully inspected, validated, or justified by ordinary means, and thus increase the risk of undetected or unforeseen errors, biases, and harms.

Personally Identifiable Data

- Banking and finance data
- Connected devices, IoT, wearables and telecommunications data
- Consumer, purchase and loyalty data
- Education data
- Employment data
- Government data
- Health data
- Immigration data
- Insurance and legal data
- Social media and content data
- Transport data

IEEE – Protection of Personally Identifiable Data

- Individuals should have access to trusted identity verification services to validate, prove, and support the context-specific use of their identity. Regulated industries and sectors such as banking, government, and telecommunications should provide data-verification services to citizens and consumers.

IEEE – Recommendations Regarding Government Use of AIS

1. Government stakeholders should identify **the types of decisions and operations that should never be delegated to AIS**, such as when to use lethal force, and adopt rules and standards that ensure effective human control over those decisions.

2. Governments should not employ AIS that cannot provide an account of the law and facts essential to decisions or risk scores.

3. AIS should be designed with transparency and accountability as primary objectives. **The logic and rules embedded in the system must be available to overseers of systems,** if possible. If, however, the system's logic or algorithm cannot be made available for inspection, then alternative ways must be available to uphold the values of transparency. Such systems should be subject to risk assessments and rigorous testing.

4. Individuals should be provided a forum to make a case for extenuating circumstances that the AIS may not appreciate — in other words, **a recourse to a human appeal.**

5. Automated systems should generate **audit trails** recording the facts and law supporting decisions and such systems should be amenable to **third-party verification** to show that the trails reflect what the system in fact did. Audit trails should include a comprehensive history of decisions made in a case, including the identity of individuals who recorded the facts and their assessment of those facts. Audit trails should detail the rules applied in every mini-decision made by the system.

IEEE Recommendations for Legal Accountability for Harm Caused by AIS

1. Designers should consider adopting an **identity tag** standard — that is, no AIS agent should be released without an identity tag to maintain a clear line of legal accountability.

2. Lawmakers and enforcers need to ensure that the implementation of AIS is not abused by businesses and entities employing the AIS to avoid liability or payment of damages. Governments should consider adopting **regulations requiring insurance** or other guarantees of financial responsibility so that victims can recover damages for harm caused by AIS.

3. Companies that use and manufacture AIS should be required to establish **written policies** governing how the AIS should be used, including the real-world applications for such AI, any preconditions for its effective use, who is qualified to use it, what training is required for operators, how to measure the performance of the AIS, and what operators and other people can expect from the AIS.

4. Because the person who activates the AIS will not always be the person who manages or oversees the AIS while it operates, states should avoid adopting universal rules that assign legal responsibility and liability to the person who "turns on" the AIS.

IEEE Recommendations for Legal Accountability for Harm Caused by AIS

5. For the avoidance of repeated or future harm, companies that use and manufacture AIS should consider the importance of **continued algorithm maintenance**. Design does not end with deployment. Thus, there should be a clear legal requirement of (1) due diligence, and (2) sufficient investment in algorithm maintenance on the part of companies that use and manufacture AIS that includes monitoring of outcomes, complaint mechanism, inspection, correction, and replacement of harm-inducing algorithm, if warranted. Companies should be prohibited from contractually delegating this responsibility to unsophisticated end-users.

6. Promote **international harmonization of laws** related to liability in the context of AIS design and operation (through bi- or multilateral agreements) to enhance interoperability, and facilitate transnational dispute resolution.

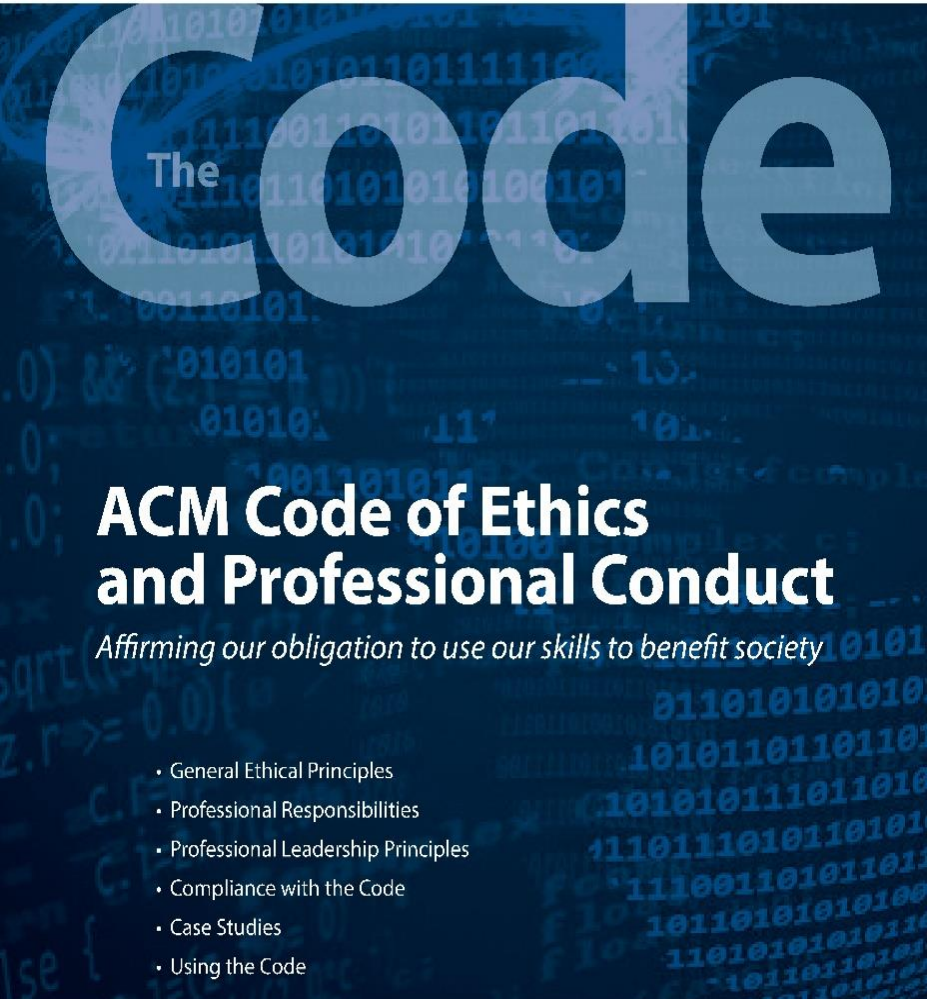
7. Courts weighing AIS litigation cases based on some form of injury should adopt a similar scheme to that of product liability litigation, wherein companies are not penalized or held responsible for installing post-harm fixes on their products designed to make the product safer.

IEEE Recommendations for Disclosure

1. **A government-approved labeling system** like the skull and crossbones found on household cleaning supplies that contain poisonous compounds could be used for this purpose to improve the chances that users are aware when they are interacting with AIS.
2. AIS should be programmed so that, under certain high risk situations where human decision-making is involved, they **proactively inform users of uncertainty** even when not asked.
3. With any significant potential risk of economic or physical harm, designers should conspicuously and adequately warn users of the risk and provide a greater scope of proactive disclosure to the user. Designers should remain mindful that some risks cannot be adequately warned against and should be avoided entirely.
4. To reduce the risk of AIS that are unreasonably dangerous or that violate the law from being marketed and produced, lawmakers should provide whistleblower incentives and protections.

Association for Computing Machinery

Advancing Computing as a Science & Profession



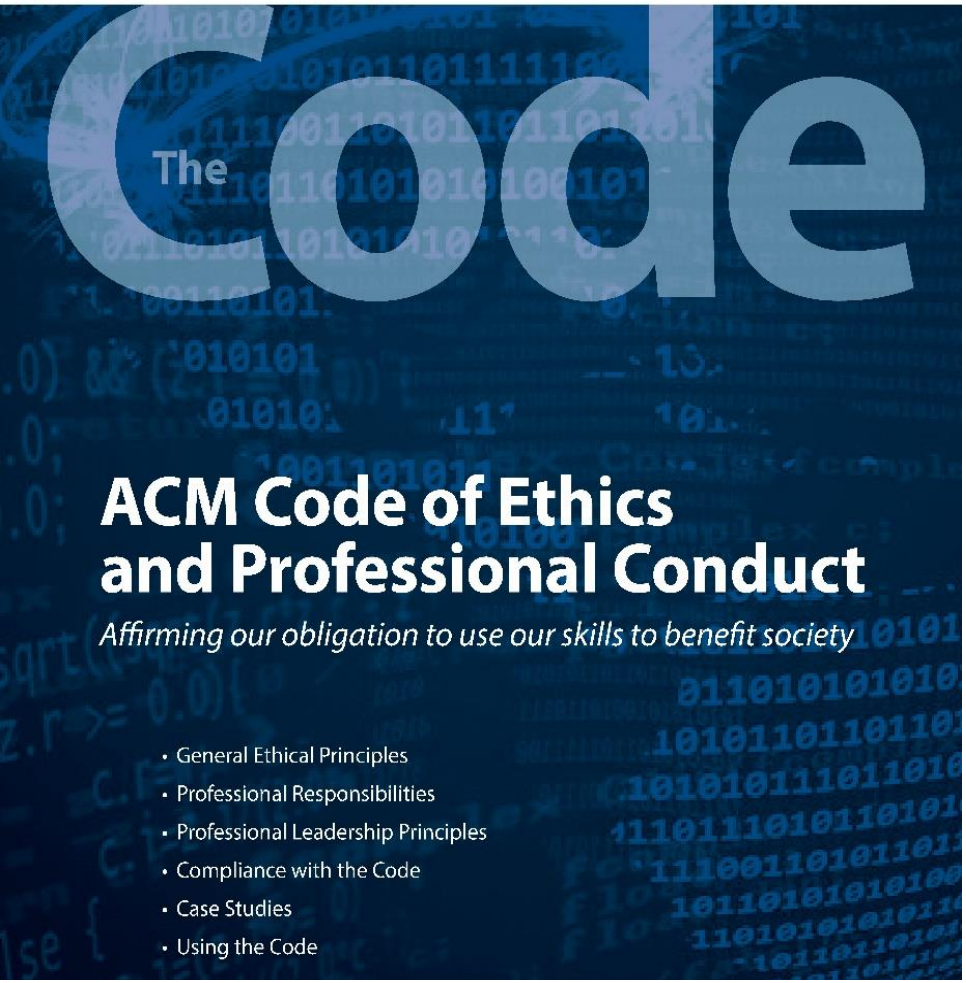
"A computing professional should be transparent and provide full disclosure of all pertinent system capabilities, limitations, and potential problems to the appropriate parties."

ethics.acm.org ACM Code of Ethics and Professional Conduct

JONES DAY

Association for Computing Machinery

Advancing Computing as a Science & Profession



ethics.acm.org ACM Code of Ethics and Professional Conduct

"The entire computing profession benefits when the ethical decision-making process is accountable to and transparent to all stakeholders. Open discussions about ethical issues promote this accountability and transparency."

JONES DAY®

Does the General Data Protection Regulation ("GDPR") Establish a Right of Explanation?

- Automated Decision-Making

  Article 22(1) states: "*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her*".

  Such processing would have to be justified on one of three bases set out as exceptions under Article 22(2), namely: performance of a contract, authorized under law, or explicit consent.

Does GDPR Establish a Right of Explanation?

Articles 13-15 cover separate aspects of a data subject's right to understand how her/his data is being used. Importantly, each article states that the data subject has the right to access "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." Articles 21-22 suggest that the subject's right to understand "meaningful information" about and the "significance" of automated processing is related to her/his right to opt out.

Recital 71 states that automated processing "should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."

Non-Compliance with GDPR Carries Heavy Fines

The GDPR establishes a tiered approach to penalties for breach which enables the DPAs to impose fines for some infringements of up to the higher of 4% of annual worldwide turnover and EUR20 million. Other specified infringements would attract a fine of up to the higher of 3% of annual worldwide turnover and EUR10m.

GDPR Applies to Companies Outside the EU

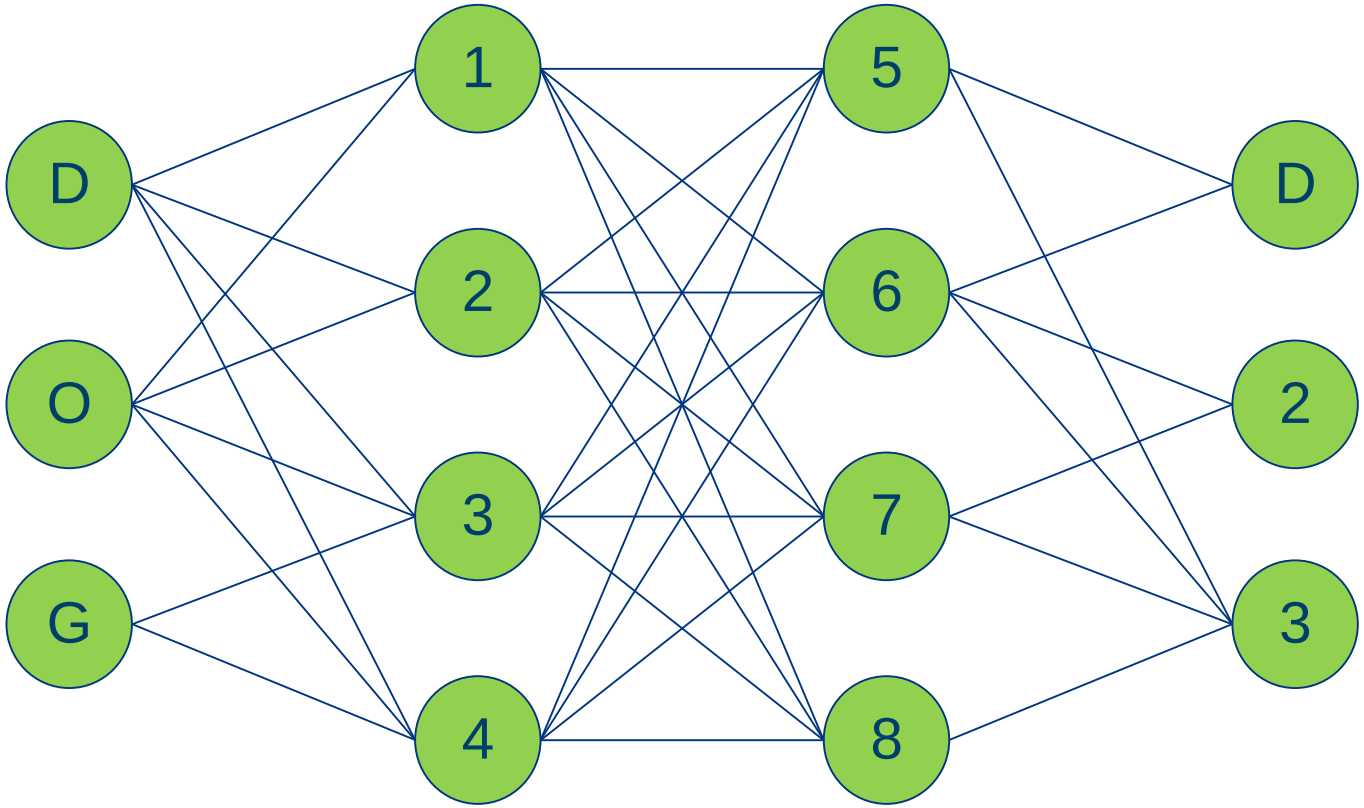
The GDPR applies to data controllers and processors outside the EU whose processing activities relate to the offering of goods or services (even if for free) to, or monitoring the behavior of, EU data subjects. Many will need to appoint a representative in the EU.

The Recitals state: "Offering goods or services" is more than mere access to a website or email address, but might be evidenced by use of language or currency generally used in one or more Member States with the possibility of ordering goods/services there, and/or mentioning customers or users who are in EU. "Monitoring of behavior" will occur, for example, where individuals are tracked on the internet by techniques which apply a profile to enable decisions to be made/predict personal preferences, etc.

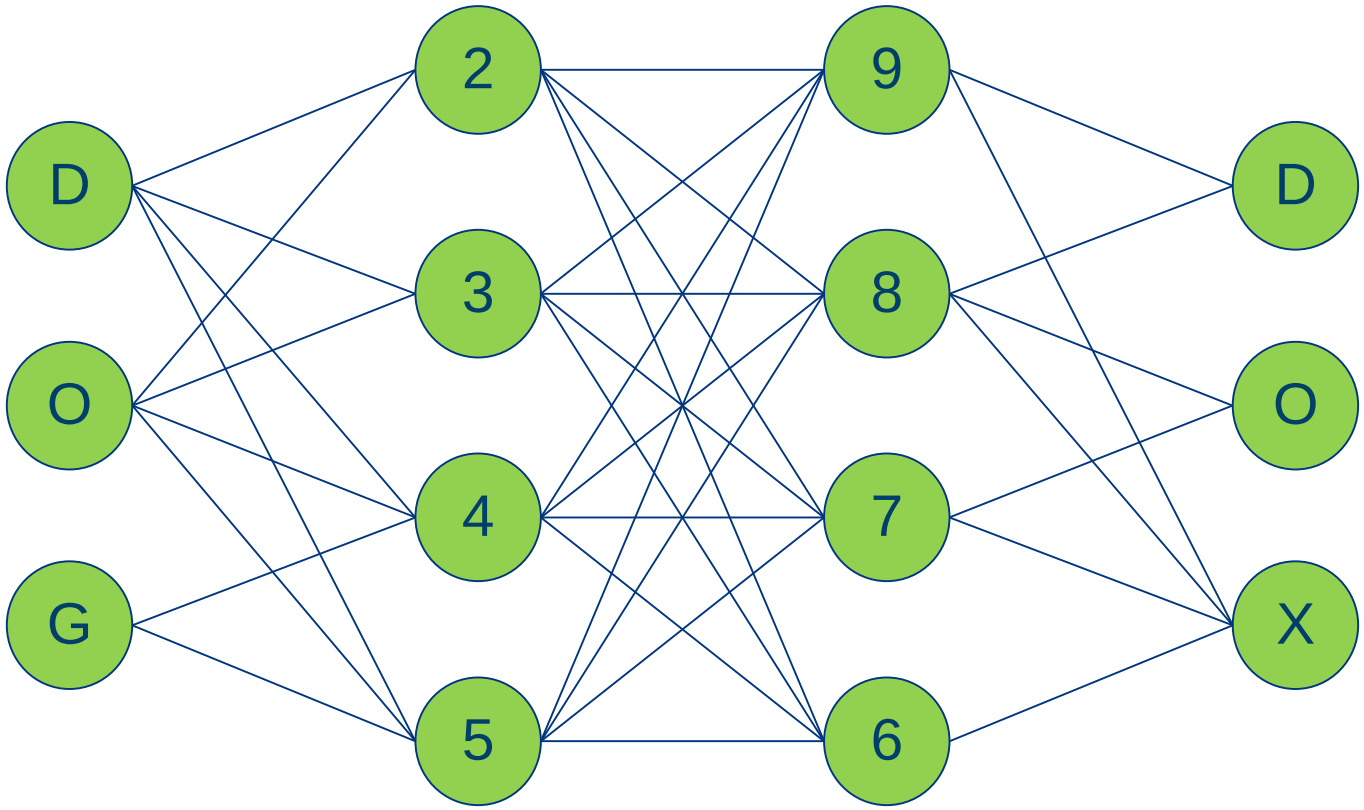
Technical Challenges to Explainable AI

- Neural networks are so called because they mimic – approximately – the structure of the brain. They are composed of a large number of processing nodes that, like individual neurons, are capable of only very simple computations but are connected to each other in dense networks.
- In a process referred to as "deep learning," training data is fed to a network's input nodes, which modify it and feed it to other nodes, which modify it and feed it to still other nodes, and so on. The values stored in the network's output nodes are then correlated with the classification category that the network is trying to learn – such as the objects in an image, or the topic of an essay.
- Over the course of the network's training, the operations performed by the individual nodes are continuously modified to yield consistently good results across the whole set of training examples. By the end of the process, the computer scientists who programmed the network often have no idea what the nodes' settings are. Even if they do, it can be very hard to translate that low-level information back into an intelligible description of the system's decision-making process.

Technical Challenges to Explainable AI



Technical Challenges to Explainable AI



Working on an Explanation

- Harvard University Berkman Klein Center Working Group on Explanation and the Law says:
  - A.I. and A.I. devices should give the reasons or justifications for a particular outcome, but not a description of the decision-making procedures.
  - Also, we need to know whether changing a factor would change the decision.
  - This explanation should be given whenever a human (or corporation) would have to give an explanation.
  - But what about the weighing of factors (i.e., judgment)?

DARPA Research

U.S. Defense Advanced Research Project Agency (“DARPA”) is working on making systems controlled by artificial intelligence more accountable to their human users.

- The Explainable Al (XAI) program aims to create a suite of machine learning techniques that:
  - Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

The New York Times

March 7, 2018

Google Researchers Say They're Learning How Machines Learn

By CADE METZ

SAN FRANCISCO — Machines are starting to learn tasks on their own. They are identifying faces, recognizing spoken words, reading medical scans and even carrying on their own conversations.

All this is done through so-called neural networks, which are complex computer algorithms that learn tasks by analyzing vast amounts of data. But these neural networks create a problem that scientists are trying to solve: It is not always easy to tell how the machines arrive at their conclusions.

On Tuesday, a team at Google took a small step toward addressing this issue with the unveiling of new research that offers the rough outlines of technology that shows how the machines are arriving at their decisions.

"Even seeing part of how a decision was made can give you a lot of insight into the possible ways it can fail," said Christopher Olah, a Google researcher.

A growing number of A.I. researchers are developing ways to better understand neural networks. Jeff Clune, a professor at University of Wyoming who now works in the A.I. lab at the ride-hailing company Uber, called this "artificial neuroscience."

Understanding how these systems work will become more important as they make decisions now made by humans, like who gets a job and how a self-driving car responds to emergencies.

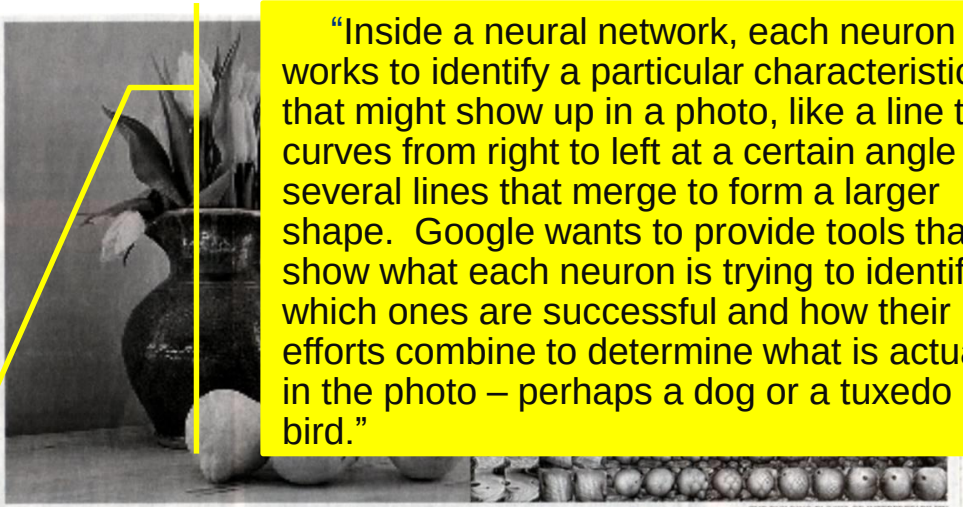
First proposed in the 1950s, neural networks are meant to mimic the web of neurons in the brain. But that is a rough analogy. These algorithms are really series of mathematical operations, and each operation represents a neuron. Google's new research aims to show — in a highly visual way — how these mathematical operations perform discrete tasks, like recognizing objects in photos.

Inside a neural network, each neuron works to identify a particular characteristic that might show up in a photo, like a line that curves from right to left at a certain angle or several lines that merge to form a larger shape. Google wants to provide tools that show what each neuron is trying to identify, which ones are successful and how their efforts combine to determine what is actually in the photo — perhaps a dog or a tuxedo or a bird.

The kind of technology Google is discussing could also help identify why a neural network is prone to mistakes and, in some cases, explain how it learned this behavior, Mr. Olah said. Other researchers, including Mr. Clune, believe they can also help minimize the threat of "adversarial examples" — where someone can potentially fool neural networks by, say, doctoring an image.

Researchers acknowledge that this work is still in its infancy. Jason Yosinski, who also works in Uber's A.I. lab, which grew out of the company's acquisition of a start-up called Geometric Intelligence, called Google's technology idea "state of art." But he warned it may never be entirely easy to understand the computer mind.

"To a certain extent, as these networks get more complicated, it is going to be fundamentally difficult to understand why they make decisions," he said. "It is kind of like trying to understand why humans make decisions."



THE BUILDING BLOCKS OF INTERPRETABILITY

On the left: an image that was put through a neural network trained to classify objects in images — for example, to tell if an image includes a vase or a lemon. On the right: a visualization of what one layer in the middle of the network detected at each position.

"Inside a neural network, each neuron works to identify a particular characteristic that might show up in a photo, like a line that curves from right to left at a certain angle or several lines that merge to form a larger shape. Google wants to provide tools that show what each neuron is trying to identify, which ones are successful and how their efforts combine to determine what is actually in the photo – perhaps a dog or a tuxedo or a bird."

JONES DAY®